学校编码：10384　　分类号_____密级____
学号：23020081153263　　UDC____

厦门大学

硕 士 学 位 论 文

基于结构和内容的 XML 文档分类的研究

Research on XML Document Classification Based on Structure and Content

张 娜

指导教师姓名：张东站 副教授
专 业 名 称：计算机应用技术
论文提交日期：
论文答辩时间：
学位授予日期：

答辩委员会主席：_____
评 阅 人：_____

2011 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

摘　要

随着网络技术的飞速发展，信息大量膨胀和聚集，互联网已经形成了一个由数据构成的巨大数据仓库，XML（eXtensible Markup Language）作为一种常用的数据交换和传输标准，蕴含了丰富的信息，具有通用的数据表示能力，能表示结构化、半结构化及元结构化的数据。因此，对 XML 文档的挖掘已经成为数据挖掘一个新的研究热点。其中，对 XML 文档分类的研究越来越广泛。根据 XML 文档的性质，分类时 XML 文档结构有许多模型，有基于树的、基于图的和基于路径的等等，其中 XML 文档的结构相似性度量是 XML 结构分析的核心问题。将 XML 文档视为一棵标记树时，已有的 XML 文档结构相似性度量主要包括距离编辑法、路径匹配法和时序分析法等。除结构以外，XML 文档的内容对 XML 文档分类的影响也很重要，所以从结构和内容两方面研究 XML 文档分类方法具有重要的理论意义和广泛的实用价值。

本文基于结构和内容两个方面对 XML 文档分类模型和算法进行了深入研究。首先，针对目前 XML 文档基于结构和内容的编辑距离分类算法的不足，本文在计算相似性度量时提出了一种新的改进方法 CS-XMLSim 方法，使得当 XML 文档结构相似而内容差异大时，分类有较高的准确率。实验结果表明，当 XML 文档结构相似内容差异大时，CS-XMLSim 方法在提高分类准确率方面有了明显改善。

其次，针对传统 KNN 算法的不足，在 CS-XMLSim 算法作为相似度计算的基础上，本文提出了 KNN 的改进算法 DB-KNN 算法。DB-KNN 算法是基于聚类和密度的 KNN 改进算法，DB-KNN 算法根据训练样本的密度采用聚类的方法，除去训练集中一定数量的噪声样本，使样本在类别内分布地更加均匀，在提高文本分类准确率的同时，减少了样本间相似度的计算量，克服了 KNN 分类过程中搜索空间巨大的问题。最后通过实验验证了 DB-KNN 算法的有效性和高效性。

关键字：XML 文档分类；相似性度量；KNN

Abstract

With the rapid development of network technology, the scale of imformation expends extremely. The Internet has already become a large data warehouse. XML (eXtensible Markup Language) includes a wealth of information as a common standard for data presentation and exchange. It has universal data processing capability and can represent structured and semi-structured data. Thus, XML document mining is becoming a new hotspot of data mining and the study of XML document classification become more and more widely. According to the properties of XML documents, the structure of XML document includes many models, such as tree-based, graph-based and path-based, etc., and the structure of XML document similarity measure is the core issues. As XML document be seen a tree, the existing XML document structure similarity measures include distance editing method, matching the path, timing analysis method and so on. Besides structure, the influence of content for classification is also important. So the study of XML document classification based on structure and content is of great significance in theory and practice.

In this article, we study on XML document categorization algorithm based on both structure and content. Considering the disadvantage of the current classification of XML documents that based on structure and content, this paper presents an improved method called CS-Similarity computing similarity measure, which maintains an high accuracy rate when XML documents are similar in structure but different in content. A better result can be seen when classifying XML documents which are more focus on content and it is more effective on classifying XML documents. The experiments prove that when XML documents are similar in structure but different in content, CS-Similarity in this paper provides a significant improvement in improving classification accuracy rate.

Secondly, considering the disadvantage of traditional KNN algorithm, this paper

presents an improved method called DB-KNN, which based on CS-XMLSim computing similarity measure. DB-KNN is an improved algorithm of KNN based on clustering and density. It reduces an amount of noisy samples based on clustering according to density of training samples in order to training samples distribution is more uniform. This approach not only enhances precision but also reduces the calculation computing document similarity, and it overcomes the difficulty that the search space in the process of KNNalgorithm is quite huge. Finally, the experiment demonstrates the effectiveness and superiority of this algorithm.

Key words: XML Document Classification; Similarity Measure; KNN

目　录

Contents

第一章 绪论

1.1 研究背景及意义

数据挖掘（Data Mining）是指从大量的数据中获取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程[1]。按照广义的观点，数据挖掘就是从存放在数据库、数据仓库或其他信息库中的大量的数据中“挖掘”有趣知识的过程。同时，数据挖掘又被称为数据库中的知识发现(Knowledge Discovery in Databases，KDD)[2]，因此，也有人把数据挖掘视为数据库知识发现过程的一个基本步骤。

数据挖掘有四种主要的任务：预测建模、聚类分析、关联分析和异常检测。其中，预测建模涉及两类预测建模任务：用于预测离散的目标变量的分类和用于预测连续目标变量的回归。这些信息的表现形式为规则、概念、规律及模式等。它可帮助决策者分析历史数据及当前数据，并从中发现隐藏的关系和模式，进而预测未来可能发生的行为[1]。数据挖掘的核心技术是人工智能、机器学习和统计学。但是，一个数据挖掘系统不是多项技术的简单组合，而是一个完整的体系，它需要辅助技术的支持。通常，它由数据库管理模块、挖掘前处理模块、挖掘操作模块、模式评估模块、知识输出模块组成。

随着网络技术的飞速发展，信息大量膨胀和聚集，互联网已经形成了一个由数据构成的巨大数据仓库，XML（eXtensible Markup Language）作为一种常用的数据交换和传输标准，蕴含了丰富的信息，具有通用的数据表示能力，能表示结构化、半结构化及元结构化的数据[3]。因此，对 XML 文档的挖掘已成为从互联网上快速有效获取信息的最佳途径之一，而且也已成为数据挖掘技术的一个新研究热点。

XML 文档挖掘，顾名思义就是对 XML 文档表示的数据进行数据挖掘。于是，XML 文档分类作为 XML 文档挖掘的一个重要组成部分，也逐渐成为国内外学者研究和讨论的热点。因此对 XML 文档的挖掘变得日益重要[4,5]，XML 文

档分类的研究也越来越广泛。

根据 XML 文档的性质，分类时 XML 文档结构有许多模型，有基于树的、基于图的和基于路径的等等。其中，XML 文档的结构相似性度量是 XML 结构分析的核心问题。将 XML 文档视为一棵标记树时，已有的 XML 文档结构相似性度量主要包括距离编辑法、路径匹配法和时序分析法等。除结构以外，XML 文档的内容对 XML 文档分类的影响也很重要，而纯文本的分类方法在 XML 文档分类上不能区分 XML 文档结构上的差异，所以从结构和内容两方面对 XML 文档分类方法的研究越来越深入。

1.2 国内外研究现状

XML 文档挖掘是一个很有研究价值，但同时也是一个比较新兴的研究课题。近年来，国内外学者从不同层面、采用不同的方法和技术对该课题进行研究，取得了不少有益的成果，提出了一些很有特色的策略和算法。但是，由于 XML 相关技术标准本身的不断修改和 XML 应用的飞速发展，目前并不存在一个成熟完整的解决方案，大部分的算法都仅仅处于理论研究阶段，它们各有优缺点，各有自己特定的使用环境，很少真正在商业化的应用系统中实施。

我们结合 XML 的特点和挖掘目标将 XML 数据分为结构挖掘和内容挖掘两大类。XML 结构挖掘就是把 XML 文档的结构当作是一棵有序、有根的标记树，对 XML 文档树进行挖掘。XML 内容挖掘就是对文档中每个开始标记和结束标记之间的文本部分进行挖掘。XML 的内容挖掘主要有三种途径[6]：第一种是利用 XML 查询语言如 XML-QL、XML-GL、XQUERY[7,8]等的查询功能，嵌入到应用程序，获得数据集进行挖掘，它将数据挖掘与 XML 紧密结合起来，但是查询开销大，修改困难；第二种是结合关系数据挖掘方法，将结构化的 XML 数据映射到现有的关系模型或对象模型中对其进行挖掘，可是 XML 的自有特点会导致在映射过程中产生问题；第三种结合文本挖掘方法，将 XML 转换成文本进行挖掘，如使用矢量空间模型(VSM)[9]将文档空间看作是由一组正交词条矢量所组成的矢量空间，通过统计词频、缩减维数等步骤，达到机器学习和获得知识的目的，但是没有考虑到 XML 文档仍然可能存在结构化特征，同时数据量大会致使